SISTEM TEMU KEMBALI INFORMASI PADA DOKUMEN DENGAN METODE VECTOR SPACE MODEL

Irmawati

Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional
e-mail : irmawati@civitas.unas.ac.id

ABSTRAK

Informasi saat ini sangat mudah didapatkan dengan memanfaatkan fasilitas internet dimanapun dan kapanpun. Di sisi lain informasi yang didapat dari search engine merupakan semua hal yang berkaitan dengan kata kunci yang dicari. Hal ini menyebabkan pengguna terpaksa menyaring untuk mendapatkan dokumen yang relevan. Oleh karena itu diperlukan cara untuk mengelompokkan banyaknya informasi yang tersedia, yang dibutuhkan pengguna sehingga memudahkan pengguna untuk mendapatkan dokumen yang diinginkan. Pada penelitian ini diusulkan suatu solusi dari permasalahan tersebut dengan mengembangkan metode ilmu pencarian yang dikenal dengan temu-kembali informasi (information retrieval) dan metode Vector Space Model (VSM). Pada metode Vector Space Model (VSM) beberapa dokumen online akan diindeks dan diurutkan berdasarkan bobot dari kata pencarian yang terdapat di dalam dokumen online tersebut. Salah satu algoritma pembobotannya adalah algoritma tf-idf yang dipengaruhi oleh frekuensi kemunculan kata pada tiap dokumen online dan frekuensi dari dokumen online yang memiliki kata tersebut.
Kata kunci: Vector Space Model, Tf-idf, Sistem Temu-Kembali Informasi, Cosine Similarity

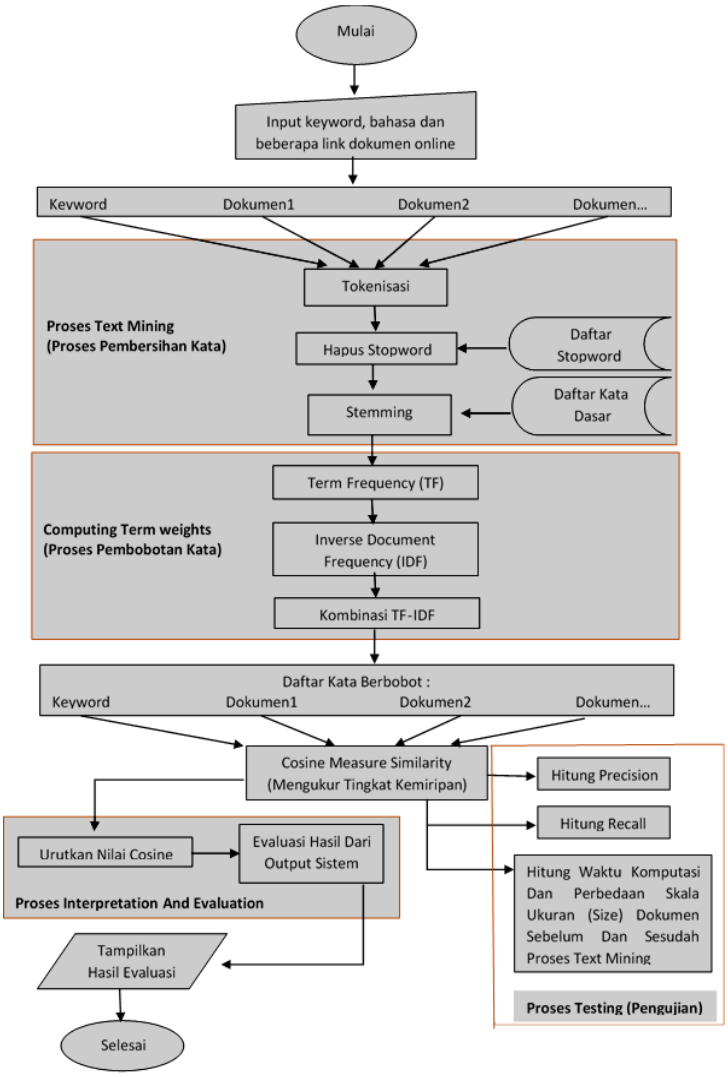
1. PENDAHULUAN

Peningkatan arus informasi yang sangat cepat dalam mendukung kegiatan *browsing dan searching* bagi *user* untuk mempermudah aktivitas mereka dalam mendapatkan informasi secara cepat, relevan, dan sesuai kebutuhan yang diinginkan. Hal ini diikuti juga dengan berkembangnya teknologi *Information Retrieval* (IR) yang merupakan sebuah sistem pencarian materi (dokumen teks) dari sifat yang tidak terstruktur (teks) sehingga mampu memenuhi kebutuhan informasi dari sekumpulan dokumen yang besar (pada *server* kompuer lokal atau *internet*). Secara prinsip, penyimpanan informasi dan penemuan kembali informasi (*information retrieval system*) adalah hal yang sederhana, misalkan terdapat tempat penyimpanan dokumen-dokumen (corpus) dan *user* merumuskan suatu pertanyaan (*request atau keyword*) yang jawabannya adalah himpunan dokumen yang mengandung informasi yang diperlukan yang diekspresikan melalui pertanyaan user. Melihat hasil penilitian yang sudah ada sebelumnya, yang terkait dalam penerapan sistem temu kembali informasi, banyak sistem yang menggunakan dokumen berbasis *offline,* yang mana saat ini banyak *user* yang beralih ke internet sebagai media pencarian informasi penting yang dibutuhkannya secara teraktual.

Penelitian sebelumnya dalam mengenai sistem temu kembali informasi juga menunjukan bahwa sistem yang dibangun tidak menghasilkan kata secara terurut melainkan secara acak terhadap hasil kemunculan *keyword* di setiap dokumen dalam *corpus*. Algoritma yang digunakan dalam proses *stemming* (mengubah kata kembali ke kata dasar) hanya didasarkan pada bahasa Indonesia saja sehingga untuk bahasa lain sistem tidak akan bekerja. Salah satu alternatif pengklasifikasian dokumen berdasarkan nilai tingkat kemiripan antara dokumen-dokumen yang ada dengan *keyword* yang dimasukan serta meningkatkan tingkat relevansi hasil *retrieval* dokumen tersebut ke dalam sebuah sistem temu kembali informasi (*Information Retrieval)* yaitu dengan menggunakan algoritma *suffix tree clustering* dan *Vector Space Model*, dibandingkan dengan metode lain dalam melakukan pengklasifikasian dokumen, metode *suffix tree clustering* dan *Vector Space Model* memiliki beberapa kelebihan, nilai pemeringkatan secara jelas dalam pengambilan informasi, penyocokan secara *partial* terhadap *keyword* dan juga menghasilkan hasil referensi yang sesuai dengan kebutuhan.

Pada metode *suffix tree clustering* dan *Vector Space Model,* setiap dokumen serta *keyword* yang telah dilakukan proses *Text Mining* kemudian diberikan bobot kata masing-masing yang terkandung disetiap dokumen yang ada dengan algoritma pembobotan kata *Term Frequency - Inverse Document Frequency* (TF-IDF). Hasil yang didapat dari pembobotan kata setiap dokumen dilakukan perhitungan pengukuran nilai tingkat kemiripan dengan membandingkan antara kedua vektor yang bersesuaian dan kemudian mengukur tingkat kemiripan terhadap *keyword* menggunakan rumus *Cosine Similarity.* Kemudian didapatkan sebuah hasil pengklasifikasian dokumen dengan tingkat kemiripan yang mendekati *keyword.*

2. METODE PENELITIAN



Gambar 1 Desain Proses Keseluruhan Sistem

Tahap-tahap dimulai dari pemrosesan data yang tidak terstruktur menjadi terstruktur sampai pada penyaringan data hingga ditemukan sebuah *knowledge* atau relevansi hasil informasi yang diperlukan *user* oleh sistem, dijabarkan sebagai berikut:

1) *User* diharuskan untuk memasukan *link* dokumen *online* (minimal 10 *link*), penggunaan bahasa (Indonesia dan Inggris) yang ingin digunakan dan *keyword*.
2) Sistem menyimpan *link* dokumen *online* dan *keyword* yang telah dimasukan oleh *user* kedalam basis data.
3) Proses *Text Mining*, pada tahap ini dilakukan teknik *Text Mining* pada dokumen *online* yang telah didapat dari langkah sebelumnya yang kemudian akan dibersihkan dan dipersiapkan untuk tahap selanjutnya. Proses untuk mempersiapkan dokumen meliputi proses pembersihan dokumen dari *tag-tag* HTML dan karakter-karakter yang tidak diperlukan, proses penghapusan *stopword* (kata penghubung) dan proses *stemming*. Proses *Text Mining* meliputi proses Tokenisasi, penghapusan *Stopword* dan proses *Stemming*.
4) *Proses Interpretation and Evaluation*, dalam proses ini pola-pola yang telah diidentifikasi oleh sistem kemudian di terjemahkan/ diintepretasikan ke dalam bentuk *knowledge* yang lebih mudah dimengerti oleh user untuk membantu dalam mengetahui hasil yang telah diberikan sistem atau bentuk lain yang lebih mudah dimengerti. Penghitungan similaritas pada langkah sebelumnya akan menghasilkan bobot pada tiap dokumen yang menentukan seberapa relevan dokumen tersebut terhadap query, sehingga dapat ditampilkan dokumen-dokumen yang relevan saja, secara terurut mulai dari yang paling relevan (bobot tertinggi).
5) Proses *Testing* (pengujian), dalam proses ini dilakukan pengujian terhadap hasil dari setiap proses yang dilakukan sistem. Untuk memperoleh perangkat lunak dengan hasil yang baik, diperlukan pengukuran kualitas terhadap hasil yang didapatkan dari sistem, yaitu:
    a) Pengujian terhadap hasil proses *Text Mining* dari sistem.
    b) Pengujian terhadap hasil proses perhitungan pembobotan kata (*Tf-Idf*) dan hasil pengukuran

tingkat kemiripan (*measure similarity*) dari sistem.

c) Pengujian terhadap waktu komputasi yang didapat dengan menggunakan *stopwatch* serta perbedaan skala ukuran (*size*) dokumen sebelum dan sesudah dilakukan proses *Text Mining* oleh sistem.

d) Pengujian terhadap hasil evaluasi yang dihasilkan sistem dengan menghitung nilai *Recall* dan nilai *Precision* untuk mengetahui tingkat optimal dari hasil sistem. Evaluasi pada sebuah sistem temu-kembali informasi dengan menggunakan *Recall* dan *Precision* sudah cukup baik untuk menjadi ukuran dari sistem tersebut [5].

e) Pengujian uji *Recall* dan *Precision* adalah untuk mendapatkan informasi hasil pencarian yang didapatkan oleh sistem temu kembali informasi yang dibuat. *Precision* dapat dianggap sebagai ukuran ketepatan atau ketelitian, sedangkan *Recall* adalah kesempurnaan. Nilai *Precision* adalah proporsi dokumen yang terambil oleh sistem adalah relevan. Nilai *Recall* adalah proporsi dokumen relevan yang terambil oleh sistem [9]. Nilai *Recall* dan *Precision* bernilai antara 0 sampai dengan 100%. Sistem temu kembali informasi diharapkan untuk dapat memberikan nilai *Recall* dan *Precision* mendekati 100% sebagai ketepatan dan kesempurnaan sistem dalam menghasilkan informasi yang relevan [9].

Sistem temu kembali informasi untuk pengklasifikasian dokumen berdasarkan tingkat kemiripan yang akan dibangun sebelumnya sudah pernah dibangun dengan tingkat keterbatasan atau kelemahan yang berbeda-beda dan juga memiliki kelebihannya masing-masing. Berikut adalah hasil analisa dari studi literatur berupa jurnal pembanding untuk mendapatkan parameter yang sudah pernah dibuat pada penelitian sebelumnya untuk bahan rumusan bagi penulis dalam membangun sistem temu kembali informasi yang lebih baik. Jurnal implementasi sistem temu kembali informasi (Information Retrieval) pada dokumen teks bahasa Indonesia menggunakan *Vector Space Model* yang ditulis oleh Taqwa Hariguna, Berlilana dan Fandy Setyo Utomo di sebuah STMIK AMIKOM Purwokerto ini bertujuan untuk membangun sebuah sistem mesin pencarian teks terhadap sebuah korpus atau database informasi berbahasa Indonesia dengan *keyword* yang spesifik yang dimasukan oleh *user*. Sistem aplikasi dibuat untuk memudahkan *user* melakukan pencarian teks kedalam korpus atau database informasi yang sesuai dan relevan terhadap *keyword* yang dimasukan *user*. Jurnal implementasi metode *Vector Space Model* pada sistem temu kembali informasi (*Information Retrieval*) yang ditulis oleh Fatkhul Amin di sebuah Universitas Stikubank pada Juli 2013 ini bertujuan untuk membangun sebuah sistem pencarian atau temu kembali informasi teks dalam sebuah korpus atau database informasi berbahasa Indonesia bedasarkan *keyword* yang dimasukan oleh *user*.

Sistem aplikasi dibuat untuk menampilkan dokumen-dokumen yang memiliki nilai kemiripan tertinggi terhadap *keyword* yang dimasukan oleh *user* dan dari hasil dokumen yang muncul dapat memiliki nilai precision yang tinggi dan nilai *recall* yang rendah untuk hasil yang relevan. Jurnal pengklasifikasian dokumen berbahasa Inggris bedasarkan *weighted-term* pada sistem temu kembali informasi (*Information Retrieval*) yang ditulis oleh Wiwin Sulistyo di sebuah Universitas Kristen Satya Wacana, Salatiga, Jawa Tengah pada Agustus 2008 ini bertujuan untuk membangun sebuah sistem pengklasifikasian dokumen teks berbahasa Inggris bedasarkan nilai *Weighted-Term* atau bobot kata pada setiap dokumen yang ada. Sistem aplikasi dibuat untuk menampilkan dokumen-dokumen yang sudah terklasifikasi dengan memiliki nilai kemiripan antara *topic* dan isi dari dokumen yang ada serta mendapatkan hasil klasifikasi dokumen secara relevan. Jurnal implementasi metode *Vector Space Model* pada sistem temu kembali informasi (*Information Retrieval*) yang ditulis oleh Oka Karmayasa dan Ida Bagus Mahendra di sebuah Jurusan Ilmu Komputer, Universitas Udayana pada Agustus 2012 ini bertujuan untuk membangun sebuah sistem pencarian atau temu kembali informasi dengan mengenal karakteristik beberapa notasi pembobotan TF-IDF serta menerapkan model ruang vektor menggunakan beberapa notasi pada metode pembobotan TF-IDF.

Sistem aplikasi dibuat untuk menemukan informasi yang relevan dengan kebutuhan dari penggunanya secara otomatis berdasarkan kesesuaian dengan *query* dari suatu koleksi informasi. Jurnal implementasi Model Ruang Vektor pada sistem informasi ruang baca di Jurusan Ilmu Komputer Universitas Udayana yang ditulis oleh Ngurah Agus Sanjaya ER, Agus Muliantara dan I Made Widiartha di sebuah Jurusan Ilmu Komputer, Universitas Udayana pada September 2012 ini bertujuan untuk membangun sebuah sistem pencarian atau temu kembali informasi teks dalam sebuah korpus atau database informasi berbahasa Indonesia bedasarkan *keyword* yang dimasukan oleh *user*. Sistem aplikasi dibuat untuk menemukan informasi yang relevan sesuai *keyword* masukan dari *user* terhadap korpus atau *database* informasi ruang baca di Jurusan Ilmu Komputer, Universitas Udayana. Sistem yang di bangun membutuhkan *keyword* untuk proses penentuan pencarian hasil terhadap data *documents source*.

3. HASIL DAN PEMBAHASAN

3.1 Modul Pengindeks

Modul pengindeks digunakan hanya pada saat sistem temu-kembali melakukan pengindeksan dokumen. Di dalam modul ini akan dilakukan *parsing*, penghilangan *stopwords* dan proses *stemming* dari dokumen-dokumen yang akan diindeks dilanjutkan dengan penghitungan nilai variabel *term frequency - invers document frequency* (*tf-idf*) dan pengukuran tingkat kemiripan (*measure similarity*). Hasil dari pengindeksan akan disimpan ke dalam basis data yang akan digunakan pada saat pencarian dokumen. Awal dari modul pengindeksan ini adalah dengan

menyimpan alamat *link* serta *title* dari tiap dokumen *online* yang telah dimasukan *user* ke dalam basis data.

Selanjutnya akan dilakukan pengambilan informasi dokumen yang akan di indeks dari basis data. Informasi yang diambil hanya didasarkan dari alamat *link* dokumen *online*, bukan dari isi dokumen *online* itu sendiri. Proses pengambilan informasi dokumen yang didasarkan dari alamat *link* dokumen *online* di dalam basis data, diawali dengan melakukan pembersihan kata dalam dokumen *online* dari karakter-karakter yang tidak dibutuhkan, seperti tanda baca, simbol, tag-tag html, javascript dan lainnya. Proses pembersihan kata ini dilakukan pada setiap dokumen *online* yang ada.

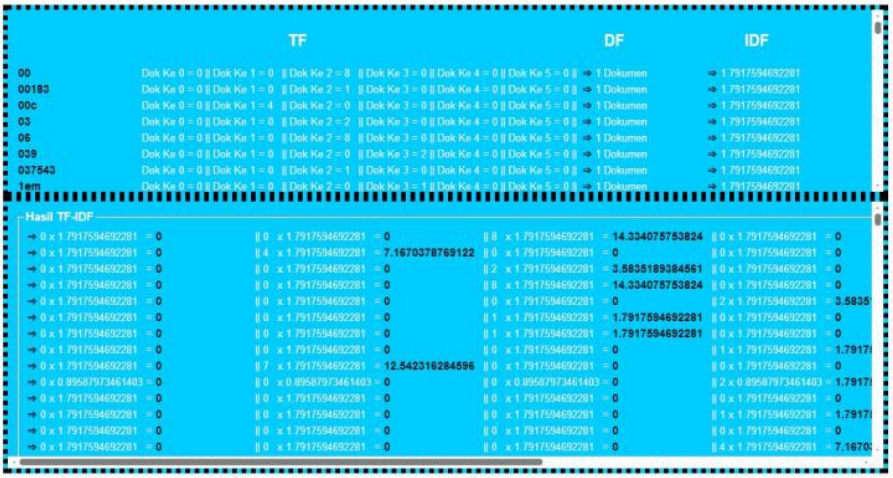
3.2 Submodul proses *Text Mining*

Setelah informasi mengenai dokumen sudah didapatkan berdasarkan hasil dari proses pembersihan kata sebelumnya, maka akan dilakukan pengecekan bahasa yang digunakan untuk temu-kembali informasi di dalam basis data yang sebelumnya telah dimasukan oleh *user*. Proses pengecekan bahasa penting dilakukan untuk memisahkan antara proses *Text Mining* dengan bahasa Indonesia dan proses *Text Mining* dengan bahasa Inggris karena masing-masing bahasa memiliki *stopwords* yang berbeda serta proses *stemming* kata yang berbeda. Hasil dari submodul ini, yaitu pengindeksan terhadap hasil dari masing-masing tahapan dalam proses *Text Mining*. Antarmuka dari halaman pengindeksan pada proses *Text Mining* tersebut dapat dilihat pada Gambar 2.

Gambar 2 Antarmuka halaman hasil proses *Text Mining*

3.3 Submodul Proses Perhitungan Pembobotan Kata (Tf-Idf)

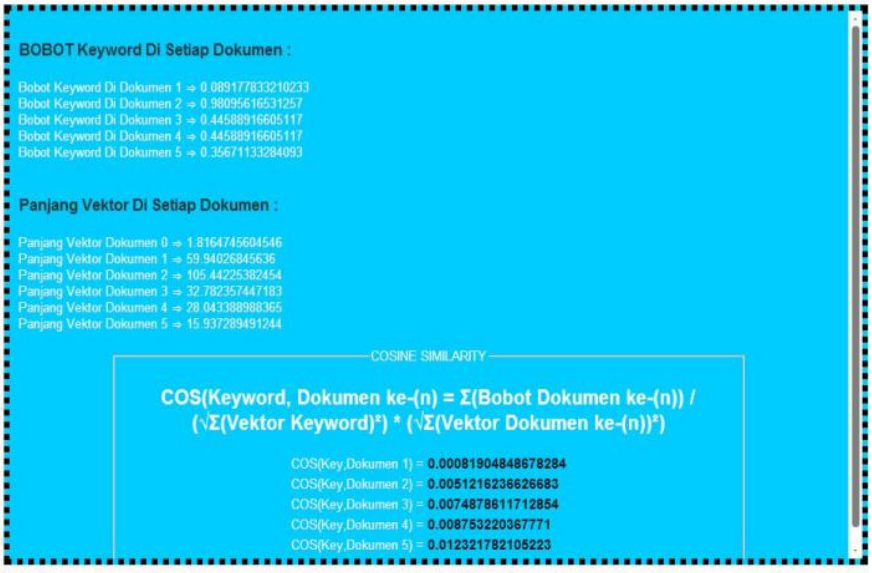
Setelah informasi telah disimpan ke dalam basis data dalam bentuk file teks, barulah informasi dapat dilakukan perhitungan pembobotan kata dengan Tf-Idf. Pembobotan kata dimulai dengan menghitung bobot setiap kata terhadap *keyword* dan seluruh informasi dalam basis data secara indeks. Pengindeksan diurutkan dimulai dari *keyword* dan kemudian informasi dari tiap basis data yang tersimpan. Hasil dari submodul ini, yaitu pengindeksan terhadap langkah-langkah serta hasil dalam proses pembobotan kata menggunakan algoritma TF-IDF. Antarmuka dari halaman pengindeksan pada proses perhitungan pembobotan kata dengan algoritma TF-IDF dapat dilihat pada Gambar 3.



Gambar 3 Antarmuka Halaman Hasil Perhitungan tf-idf

3.4 Submodul pengukuran tingkat kemiripan (*measure similarity*)

Langkah terakhir dalam modul pengindeksan apabila pada proses perhitungan pembobotan kata telah selesai dilakukan, yaitu dengan menghitung atau mengukur tingkat kemiripan vektor (isi informasi) *keyword* dengan setiap informasi dari tiap basis data. Dengan langkah awal, menghitung hasil perkalian skalar antara nilai bobot *keyword* dan setiap bobot informasi dari tiap basis data lainnya. Hasilnya perkalian dari setiap bobot informasi dengan bobot *keyword* dijumlahkan. Pada langkah diatas setiap hasil perhitungan yang didapat baik hasil perhitungan perkalian skalar maupun perhitungan panjang seluruh bobot informasi akan disimpan kedalam basis data. Selanjutnya yaitu menghitung nilai kemiripan antar vektor (informasi), dari nilai vektor *keyword* dengan nilai vektor tiap informasi sesuai urutan dalam basis data yang tersimpan. Dilakukan pemanggilan hasil dari perhitungan sebelumnya yang tersimpan didalam basis data yang dibutuhkan untuk perhitungan nilai kemiripan antar vektor. Hasil dari submodul ini, yaitu pengindeksan terhadap hasil dari pengukuran tingkat kemiripan informasi dengan rumus *Cosine Similarity* pada setiap dokumen dalam basis data pada informasi *keyword*. Antarmuka dari halaman pengindeksan terhadap hasil pengukuran tingkat kemiripan dengan rumus *Cosine Similarity* dapat dilihat pada Gambar 4.



Gambar 4 Antarmuka halaman hasil pengukuran tingkat kemiripan *(Cosinie Similarity)*

3.5 Pengujian Sistem

Pengujian dilakukan dengan menggunakan dataset berupa 5 *link* dokumen *online* yang diambil dari proses pencarian pada mesin pencari www.google.com. Dari ke 5 *link* dokumen yang didapat, masing-masing memiliki *size* yang berbeda-beda. Contoh dataset yang digunakan pada skenario pengujian ditunjukan pada Tabel 1.

Tabel 1 Dataset yang digunakan dalam pengujian

No	Keyword	Link	Title	Size (Source)
1	Pengertian Akademik	http://faiqzhahirin.blogspot.com/2013/02/pengertian-prestasi-akademik-prestasi.html	MUH. FAIQ ZHAHIRIN: Pengertian Prestasi Akademik, Prestasi Belajar , dan Prestasi Kerja	376 kb
2		http://indo-dinamis.blogspot.com/2013/04/kualifikasi-akademik-kompetensi-guru.html	Notes of My Life	238 kb

No	Keyword	Link	Title	Size (Source)
3		http://id.shvoong.com/social-sciences/education/2090542-pengertian-akademi/	Pengertian Akademik	565 kb
4		http://tipstrategi.wordpress.com/2010/05/05/pengertian-sistem-informasi-akademik/	Pengertian Sistem Informasi Akademik | Tipstrategi's Blog	211 kb
5		http://www.poltekkesjakarta3.ac.id/?q=node/20	Pengertian dalam Akademik | Politeknik Kesehatan Kemenkes Jakarta III	69 kb

Hasil pengujian pada skenario tahap pertama dan tahap kedua menunjukan bahwa kerangka kerja dari sistem yang diusulkan dapat bekerja dengan baik ketika melakukan proses *Text Mining*. Dimulai dari hasil pengujian pada skenario tahap pertama yang menunjukan kinerja dari sistem yang diusulkan dapat memperkecil ukuran informasi dokumen *online* dari sifat yang tidak terstruktur dengan ukuran yang besar menjadi lebih terstruktur dengan ukuran yang lebih kecil.

Dengan sistem yang mampu membuat *size* pada tiap informasi dokumen *online* menjadi lebih kecil, membuat pemakaian basis data lebih hemat sebagai media penyimpanan informasi-informasi dokumen *online*. Sistem dapat bekerja secara maksimal dan efektif untuk menghasilkan informasi sesuai tujuan.

4. KESIMPULAN

Berdasarkan hasil penelitian yang berupa pengembangan sistem temu-kembali informasi atau sistem pencarian dengan metode *Vector Space Model* dapat diambil beberapa kesimpulan yaitu:

1) Proses pengindeksan dan pembobotan kata dari isi informasi dokumen *online* didalam sistem temu-kembali informasi yang dikembangkan melalui beberapa tahapan pemrosesan teks (proses *Text Mining*) yaitu Tokenisasi, penghilangan *Stopwords* dan *Stemming* kata dapat ditampilkan dalam bentuk indeks secara terstruktur.
2) Kecepatan dari proses *Text Mining* yang dilakukan pada sebuah dokumen *online* sangat berpengaruh besar pada kecepatan koneksi *internet* dan *size* dari *source* dokumen *online* tersebut.
3) Tinggi atau rendahnya nilai *Precision* yang didapat sangat berpengaruh pada banyaknya dokumen *online* yang relevan terhadap *keyword* yang dimasukan pengguna. Lebih banyak dokumen *online* yang relevan terhadap *keyword* yang dimasukan, maka semakin besar nilai *Precision* yang didapat.
4) Secara kinerja, sistem temu-kembali yang dikembangkan sudah cukup baik karena dengan rata-rata nilai *Recall* hampir 100% dan rata-rata dari nilai *Precision* sekitar 73,6% yang didapatkan, sehingga 73,6% dokumen *online* yang berhasil ditemu-kembalikan relevan dengan *keyword* yang diberikan.

DAFTAR PUSTAKA

[1] Arifin, A. Z. dan Setiono, A. N. Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma *Single Pass Clustering*. *Proceeding of Seminar on Intelligent Technology and Its Applications* (SITIA). Surabaya : ITS

[2] *English Stop Words* diambil dari *Journal of Machine Learning Research*. http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop pada 3 Februari 2014 di jam 10.00.

[3] Fatkhul Amin. (2013). Sistem Temu Kembali Informasi dengan Pemeringkatan Metode *Vector Space Model*, Jurnal Teknologi Informasi DINAMIK, Volume 18. No.2.

[4] Jajang Machpudin Suryana. (2013). Implementasi *Suffix Tree Clustering* Untuk Pengelompokan Dokumen Hasil Pencarian *Online* Pada Mesin Pencari Dan Jejaring Sosial. Universitas Komputer Indonesia.

[5] Mizzaro, S., 1998. *How Many Relevances in Information Retrieval*?, *Department of Mathematics and Computer Science University of Udine*, http://www/dimi.uniud.it/~mizzaro [25 November 2013].

[6] Ngurah Agus Sanjaya ER, Agus Muliantara dan Made Widiartha. (2012). Peningkatan Relevansi Hasil Pencarian Kata Kunci Dengan Penerapan Model Ruang Vektor Pada Sistem Informasi Ruang Baca Di Jurusan Ilmu Komputer Universitas Udayana. Jurnal Ilmu Komputer, Volume 5. No. 2.

[7] Oka Karmayasa dan Ida Bagus Mahendra. (2012). Implementasi *Vector Space Model* Dan Beberapa Notasi Metode *Term Frequency Inverse Document Frequency* (Tf-Idf) Pada Sistem Temu Kembali Informasi. Jurusan Ilmu Komputer Universitas Udayana.

[8] Robertson, Stephen. 2005. *Understanding Inverse Document Frequency: On theoretical arguments for IDF, England : Journal of Documentation,* Vol. 60, pp. 502–520.

[9] Salton, G., (1989). *Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer. Addison – Wesly Publishing Company*, Inc. USA.

[10] Salton Gerard and Christopher Buckley. (1998). *Term-Weighting Approaches In Automatic Text Retrieval. Information Processing & Management* Vol. 24, No. 5, pp. 513-523, 1988.

[11] *Tala F. Z. (2004). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands*.

[12] Taqwa Hariguna, Berlilana dan Fandy SU. *Implementation Of Information Retrieval Indonesian Text Documents Using The Vector Space Model*, STMIK AMIKOM Purwokerto.

[13] Wiwin Sulistyo. (2008). Klasifikasi Dokumen Berbahasa Inggris Berdasarkan *Weighted-Term.* Jurnal Teknologi Informasi-Aiti, Vol. 6. No. 2.

[14] Yates, R.B, (1999). *Modern Information Retrieval, Addison Wesley-Pearson international edition*, Boston. USA.